Cross-Modal Contrastive Learning for Text-to-Image Generation

Year: 2021 | Citations: 412 | Authors: Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, Yinfei Yang

Abstract

The output of text-to-image synthesis systems should be coherent, clear, photo-realistic scenes with high semantic fidelity to their conditioned text descriptions. Our Cross-Modal Contrastive Generative Adversarial Network (XMC-GAN) addresses this challenge by maximizing the mutual information between image and text. It does this via multiple contrastive losses which capture inter-modality and intra-modality correspondences. XMC-GAN uses an attentional self-modulation generator, which enforces strong text-image correspondence, and a contrastive discriminator, which acts as a critic as well as a feature encoder for contrastive learning. The quality of XMC-GAN’s output is a major step up from previous models, as we show on three challenging datasets. On MS-COCO, not only does XMC-GAN improve state-of-the-art FID from 24.70 to 9.33, but– more importantly–people prefer XMC-GAN by 77.3% for image quality and 74.1% for image-text alignment, compared to three other recent models. XMC-GAN also generalizes to the challenging Localized Narratives dataset (which has longer, more detailed descriptions), improving state-of-the-art FID from 48.70 to 14.12. Lastly, we train and evaluate XMC-GAN on the challenging Open Images data, establishing a strong benchmark FID score of 26.91.